RAID-S Technical Overview: Raid 4 and 5-Compliant Hardware and Software Functionality Improves Data Availability Through Use of XOR-Capable Disks in an Integrated Cached Disk Array

Brett Quinn
EMC Corporation
Hopkinton, MA 01748-9103
Internet: quinn_brett@isus.emc.com
Web Page: www.emc.com
Telephone: +1-508-435-1000
Fax: +1-508-435-8903

1. Introduction

1.1 Objective and Scope

The purpose of this paper is to provide a technical description of RAID-S. It is intended to give the reader an understanding of how RAID-S is architected and implemented in the EMC Symmetrix 3000/5000 series Integrated Cached Disk Array. Topics include a RAID-S taxonomy, configuration considerations, operational characteristics, performance, and implementation guidelines.

It should be noted that the RAID Advisory Board granted EMC's petition to use the conformance logo for RAID Levels one, four, and five for the Symmetrix series of ICDAs. Use of the conformance logo for RAID levels four and five were also granted for the Extended On-line Storage ICDAs in June 1996. Symmetrix is considered RAID Level one-conformant when configured with mirrored devices, RAID Level four-conformant when RAID-S is configured without Hyper-Volume Extension, and RAID Level five-conformant when RAID-S is configured with Hyper-Volume Extension.

The Symmetrix series of Intelligent Cached Disk Arrays represent a family of information storage and retrieval systems available in a broad range of capacities to address current and future business and scientific requirements. Systems provide instant and dependable access to mainframe and open platforms. For further details refer to EMC's web page.

1.2 What is RAID-S?

1.2.1 Improving data availability

RAID-S (Redundant Array of Independent Disks-Symmetrix) is a combination of hardware and software functionality that improves data availability in Symmetrix 3000 and 5000 series ICDAs by using a portion of the array to store redundancy information. This redundancy information, called *parity*, can be used to regenerate data should the data on a disk drive become unavailable.

1.2.2 Flexible availability options

RAID-S is the newest RAID solution to be delivered for the Symmetrix ICDA. RAID-1, also called *Mirroring*, was first delivered in 1991. Compared to a mirrored Symmetrix, RAID-S offers EMC users more usable capacity than a mirrored system containing the same number of disk drives. Also, with the introduction of RAID-S, users can now select the level of protection they desire for data stored in the Symmetrix. Within the same Symmetrix system, data can be protected via RAID-S, Mirroring, SRDF, and/or Dynamic Sparing.

1.2.3 Technological innovation

RAID-S employs the same technique for generating parity information as many other commercially available RAID solutions, i.e., the Boolean operation *EXCLUSIVE OR* (XOR)[i]. However, EMC is the first vendor to reduce the overhead associated with parity computation by moving the operation from controller microcode to the hardware on the disk drive itself. This is done through the use of XOR-capable disk drives. This also positions RAID-S to benefit from future improvements in internal disk subsystem communications protocol performance when SCSI is supplanted by fiber channel technology.

1.2.4 Prerequisites

RAID-S is transparent to the host operating system. The prerequisites required for RAID-S are a 3000/5000 series Symmetrix with XOR capable disk drives and the appropriate Symmetrix microcode level.

2. RAID-S Taxonomy

Like most Symmetrix features, RAID-S introduces new terminology and concepts that need to be clearly understood to properly describe the functions and components of

RAID-S. Figures 1 and 2, will be referenced in the following discussion of RAID-S terms.

2.1 Group

A RAID-S group is the set of four or eight (EOS systems only, see section 2.7) physical disks within a Symmetrix system that are related to each other for parity protection. Current implementation requires that all members of a RAID-S group must be attached to the same disk director. Figures 1 & 2 both depict RAID-S groups of four physical devices each. Note that each of the four disks are on a different Disk Director SCSI bus.

2.2 Logical Volume

A logical volume is a unit of storage implemented on a single Symmetrix disk drive. When Hyper-Volume Extension (HVE) is not used, the size of a logical volume is usually the same as a physical volume. With HVE, up to eight logical volumes can exist on a physical volume.

2.3 Rank

A rank is the set of logical volumes related to each other for parity protection. Each RAID-S group supports a minimum of one rank, and with HVE enabled, a maximum of eight ranks. Figure 2 shows a RAID-S group consisting of four 9 GB drives with four ranks defined across the group. A rank is the “horizontal layer” of logical volumes and utilizes all four SCSI paths attached to a disk director.

A rank is equivalent to a “redundancy group stripe” as defined by the RAID Advisory Board.

2.4 Data Volume

A data volume is similar to a traditional logical volume in Symmetrix terminology. It is the “virtual volume” image presented to the host operating system and defined as a separate unit address to the host. All data volumes within a rank must be the same size. There can be a maximum of 512 data volumes in a Symmetrix.

It is important to note that RAID-S **does not** “stripe” data across members of a rank as is done in traditional RAID implementations. Each data volume emulates either a complete 3380 or 3390 device or a complete FBA logical volume mapped to an Open Systems host. This is a key differentiator because it allows the group to sustain the loss of more than one member and still service requests from all the surviving members. In RAID 4/5 implementations which stripe data, the loss of more than one member would result in data loss for the entire group.

This "direct" mapping of disk images to disk drives also allows standard performance and tuning techniques to be used to manage the volumes in the rank.

2.5 Parity Volume

A parity volume is a logical volume which holds the parity information for the rank. It must be the same size as the data volumes it supports. Parity volumes do not have unit addresses and are transparent to the host software. As is true with M2[ii] volumes in a mirrored Symmetrix, parity volumes are not included in the 512 device limit within a single Symmetrix system. In fact, the parity volume is referred to as an "M2" volume and is associated with three "M1" data volumes in a 3:1 rank. This is illustrated in figure 1.

When using HVE, parity volumes are distributed amongst the members of a RAID-S group, as shown in figure 2. This **distributed parity** provides for improved performance over a single physical volume which could become a performance bottleneck in a heavy write workload.

2.6 Modes of Operation

2.6.1 Normal Mode

When a RAID-S rank is operating with all members functioning it is said to be operating in normal mode.

2.6.2 Reduced Mode

When a RAID-S rank is operating with one failed *data* volume it is said to be running in **reduced** mode. Parity protection is suspended for the rank. Referring to figure 1, the failure of device 00 would force the rank to operate in reduced mode. In figure 2, the failure of device 00 would cause the first three ranks to operate in reduced mode.

2.6.3 Non-RAID Mode

When a RAID-S rank is operating with one failed *parity* volume it is said to be running in non-RAID mode. As in reduced mode, parity protection is suspended for the rank. Again referring to figure 2, the failure of device 00 would cause the fourth rank to operate in non-RAID mode.

2.6.4 Regeneration

When a data volume fails, the data on that volume is reconstructed by XORing the parity volume with the remaining data volumes in the same rank. This process is called **regeneration** and is used in place of the normal READ command when one data volume has failed. The regenerated data is placed on the parity volume of the rank. Any subsequent request for the data will be serviced by the parity volume, which is acting as a data volume for the regenerated data.

Referring to figure 1, if device 01 were to fail, the data on volume B would be regenerated by computing the exclusive OR of the data on volumes A, C, and the parity volume.

2.6.5 Rebuild

When a parity volume fails, RAID protection is suspended for the rank. When the failed device is replaced as part of a service action, the parity volume is reconstructed. This process is called **rebuild**.

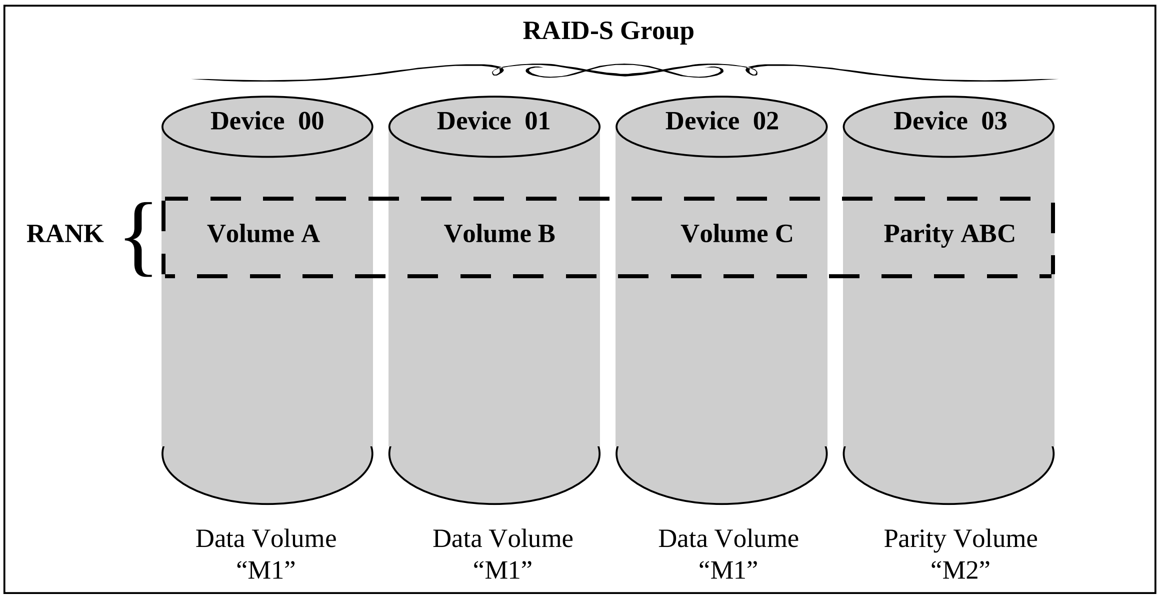

Figure 1: RAID-S Group w/o Hyper Volume

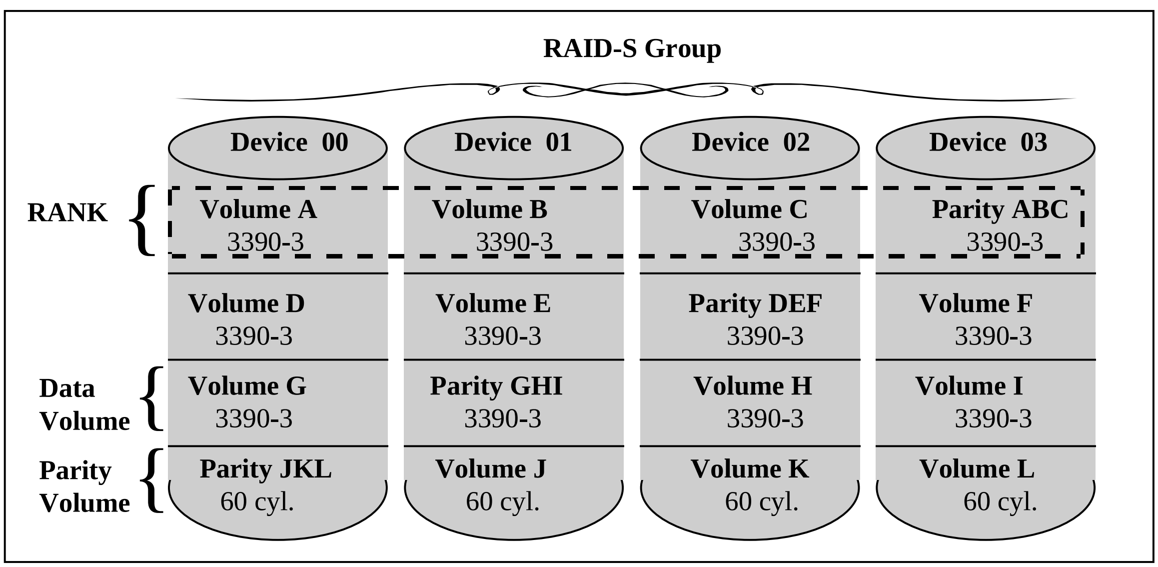

Figure 2: RAID-S Group w/Hyper Volume

2.7 EMC Extended On-line Storage (EOS)

In March 1996, the flexibility of RAID-S design and MOSAIC 2000 architecture was demonstrated with the announcement of the EOS base product (model EOS-90XX). EOS is a high capacity storage solution intended for archived data that is typically accessed in a read only mode, and where high performance is not a requirement. An EOS disk storage array offers either Dynamic Sparing or RAID-S protection for the disks in the system. The group size for EOS systems was expanded from 4 disks (3 data + 1 parity) to 8 disks (7 data + 1 parity). This has the effect of increasing the amount of storage available for user data from 75% of the array's capacity to 87.5%.

In July 1996, the EOS product line was expanded with the introduction of the EOS 9R models (EOS-9RXX). EOS 9R models offer improved performance over the base EOS models and support some of the advanced microcode features of the Symmetrix.

In both the EOS base and EOS 9R models, the number of data volumes in a rank was increased from 3 to 7. A 7+1 RAID-S group is depicted in figure 3 below. (Note that each SCSI bus now contains two members of a RAID-S group)

RAID-S in EOS systems, as in Symmetrix, can be implemented as either RAID level 4 or RAID level 5 as defined by the RAID Advisory Board. When implemented without Hyper-Volume Extension it conforms to the definition of RAID level 4. When implemented with Hyper-Volume Extension, as in figure 3, it qualifies as a RAID level 5 array.

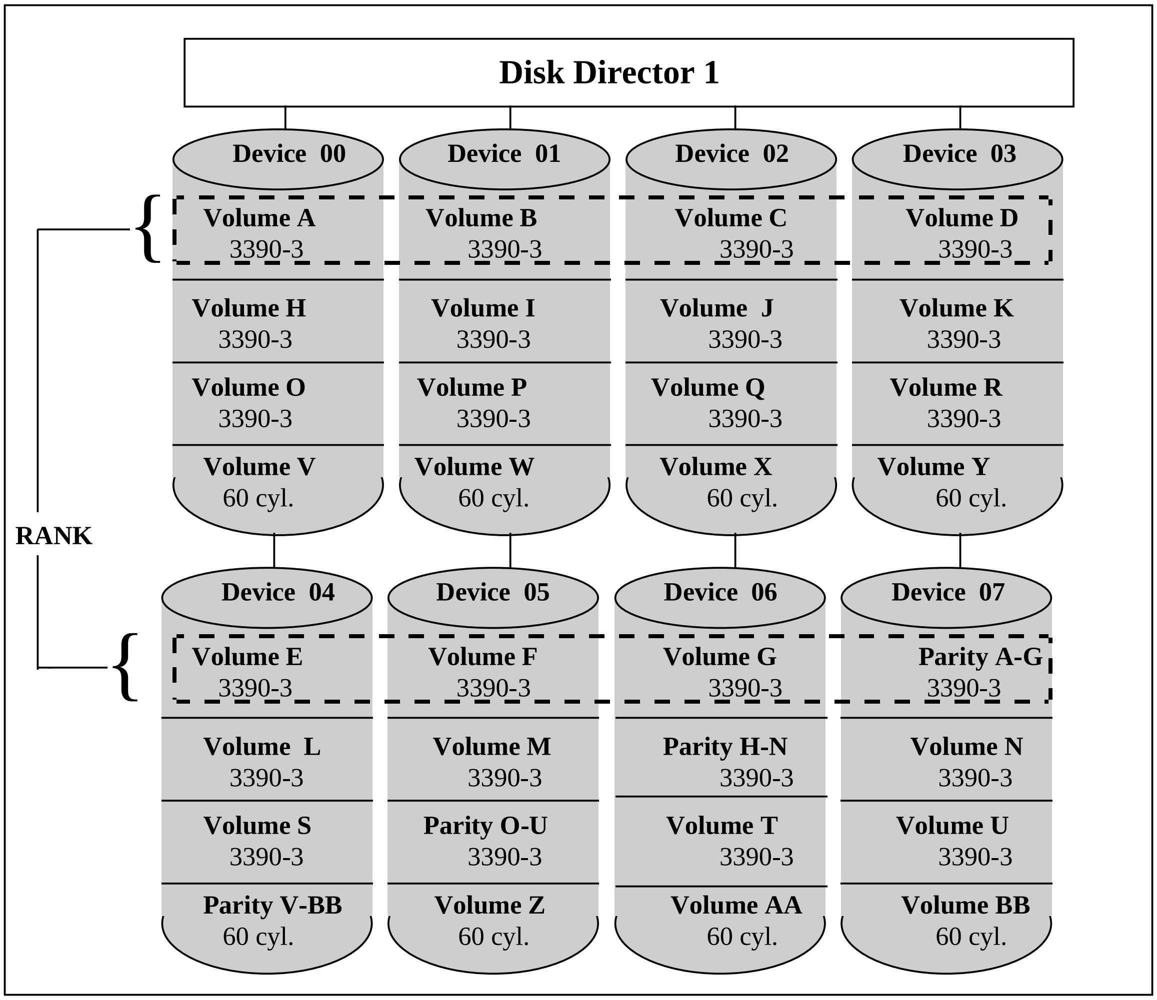

Figure 3 RAID-S 7+1 Group size in Extended On-line Storage (EOS) systems

3. Configuring RAID-S

3.1 Host addressable volumes

In a RAID-S Symmetrix, only data volumes are host addressable. Consequently, the number of host addresses is less than the number of logical volumes defined in the Symmetrix. Using a 5100-9016 (with all volumes RAID-S protected) as an example, the number of host addressable volumes is computed as follows: (each device contains 3 logical volumes, rank size is 4 volumes, 3 data :1 parity)

16 devices X 3 logical volumes/device = 48 logical volumes
25% of volumes are parity (3:1) = 12 parity volumes
48-12 = 36 host addressable volumes

3.2 Ranks and SCSI buses

Normally RAID-S configurations will have a rank size of four, with three data volumes and one parity volume per rank. In these configurations each member of the rank will be on a different SCSI bus behind the same disk director. This improves the performance of the rank by reducing SCSI bus contention during XOR calculations. The only exception to this is EOS systems which support 7+1 ranks. These implementations support two members of a group per SCSI bus.

3.3 HVE considerations

During installation and configuration of the Symmetrix 3000/5000, parity volumes are distributed across all devices in the RAID group. Obviously, the maximum number of logical volumes that can be defined on each physical device, without having two parity volumes on one device, is four. However, the maximum number of hyper-volumes allowed, including parity volumes, remains eight.

3.4 Intermixing with Local Mirroring

RAID-S groups can coexist with mirrored pairs in the same Symmetrix. It is important to remember that RAID-S groups must be defined behind the *same* disk director, while mirrored pairs must be defined behind *different* disk directors. In addition, RAID-S volumes cannot be locally mirrored, and locally mirrored volumes cannot be part of a RAID-S group.

It is possible to dynamically reconfigure a mirrored configuration to RAID-S and vice versa.

3.5 SRDF and SDM Support

RAID-S is supported with the Symmetrix Remote Data Facility and the Symmetrix Data migrator. This support is described below.

3.5.1 Symmetrix Remote Data Facility

SRDF provides the capability to remotely mirror logical volumes to another Symmetrix system. This logical volume approach is maintained in a RAID-S environment. SRDF **does not** require that a RAID-S rank or group be remotely mirrored in its entirety. Rather, SRDF simply allows a logical data volume in a rank to be remotely mirrored to another system where it can be protected via local mirroring, RAID-S, and/or dynamic sparing. Note that parity volumes are not remotely mirrored. SRDF views this remote copy (target volume) of the data as a third copy which can be accessed via the SRDF link in the event that the local copy (source volume) becomes unavailable.

This offers the benefit of using the remote copy of a volume to access data in the event the local copy is unavailable, thus avoiding the overhead of RAID-S regeneration when accessing a failed volume.

3.5.2 Symmetrix Data Migrator

SDM is a Symmetrix microcode based product which allows the direct migration of data from an existing, installed control unit (called the “donor”) to a Symmetrix (called the “target”). During a migration, the target Symmetrix is connected to a mainframe host and the donor control unit is connected to the Symmetrix. Data is then migrated from the donor to the target Symmetrix in either an on-line or off-line fashion. Parity computation can be performed during migration (the default), or after all data has been migrated to the data volumes in group.

The introduction of RAID-S protected target volumes into an SDM migration does not impact the configurability of the target Symmetrix. Donor control unit volumes are mapped to target Symmetrix volumes just as they were in a mirrored scenario.

4. RAID-S Operational Characteristics

4.1 Normal Mode operation

4.1.1 Write Operations

Fast write: As with all Symmetrix operating modes, 100% of writes are fast writes and are satisfied in the cache.

Destaging write: Write operations to a RAID-S rank are completed using a Read-Modify-Write sequence of I/O operations as depicted in figure 4 and described below.

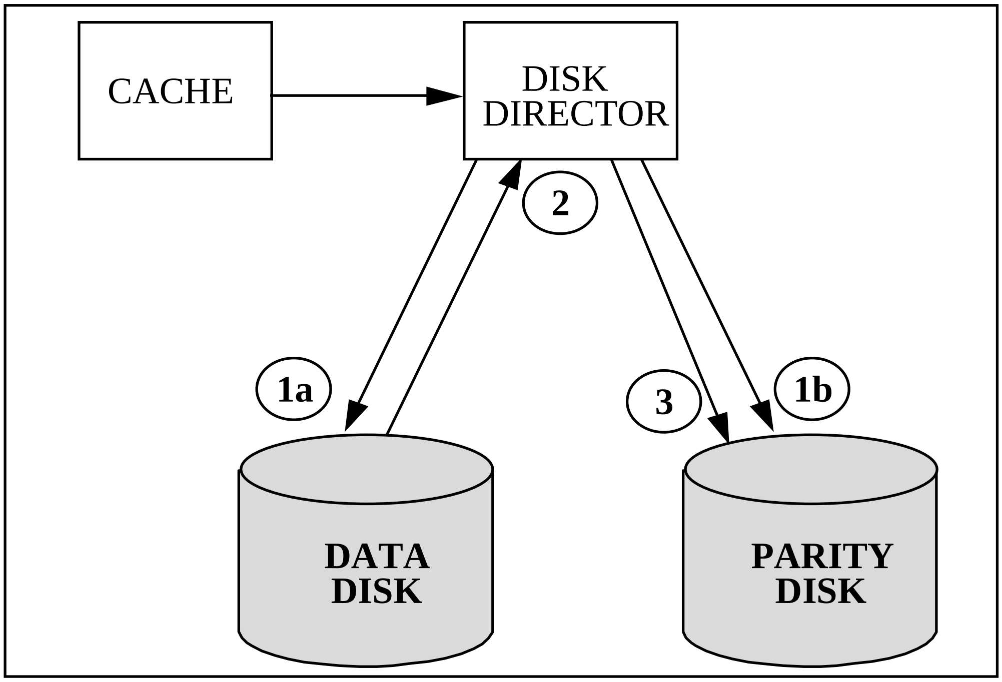

Figure 4: Read-Modify-Write Sequence

1a	The Disk Director begins the Read-Modify-Write sequence by sending the new data to the data drive using a new SCSI command called an XOR READ. This command reads the old data into the disk's buffer, XOR's it with the new data (creating difference data), and writes the new data in the disk.
1b	Simultaneously, the DD sends the parity drive another new command called an XOR WRITE (command phase only). This command instructs the parity drive to read the old parity into its buffer in preparation for XORing with the difference data from 1a.
2	The difference data is sent to the DD for transfer to the parity drive.
3	The DD sends the difference data to the parity drive during the data phase of the previously issued XOR WRITE command. The difference data is XOR'd with the old parity waiting in the buffer, and the resulting new parity is immediately written to the disk.

This Read-Modify-Write sequence constitutes the "write penalty" in RAID-S. It is significantly different from the write penalty in other RAID 4/5 implementations. The typical RAID 4/5 approach requires four discrete, sequential I/O operations be executed by the controller:

1. Read old data
2. Read old parity
3. Write new data

4. Write new parity

In addition, two processing steps must be executed by the controller microcode:

5. XOR old data with new data (creating difference data)
6. XOR difference data with old parity (creating new parity)

In contrast, RAID-S requires only two discrete sequential I/O operations be executed by the controller.

1. Write new data
2. Write difference data

The design of RAID-S distributes the work of computing parity between the disk director and the disk drives, using the XOR chip and the disk level buffer. The disk containing the data volume performs the read of the old data, the XOR to compute difference data, and sends the difference data to the disk director. The disk containing the parity volume reads the old parity (at the same time that the data drive is reading the old data), XOR's it with the difference data received from the controller, and writes the new parity to the disk.

The **parallelism** introduced into the parity computation process through the use of XOR drives allows the "controller" (disk director) to do only half the number of back-end I/Os as competitive RAID solutions. This reduces the impact of the write penalty significantly and improves the overall performance of RAID-S compared to competitive implementations.

4.1.2 Read Operations

Read hits: Read hits are processed via the cache as in normal Symmetrix processing.

Read Misses: Read misses are directed to the disk drive and processed as normal Symmetrix read misses. There is no XORing of the data, and only one disk drive is involved in servicing the request. This is a significant advantage over other RAID 4/5 implementations that "stripe" data across multiple disk drives. In these implementations more than one disk drive may be required to service the request.

4.2 Reduced Mode Operations

Note: In reduced mode operations parity protection is suspended for the rank. **No new parity data is written.**

4.2.1 Failed Data Volume

Read Miss Operation: Read requests not satisfied in the cache are called read misses, and are serviced by the disk drive. When read requests are made to a failed member of a rank the data must be regenerated to service the request. The regeneration process is depicted in figure 5.

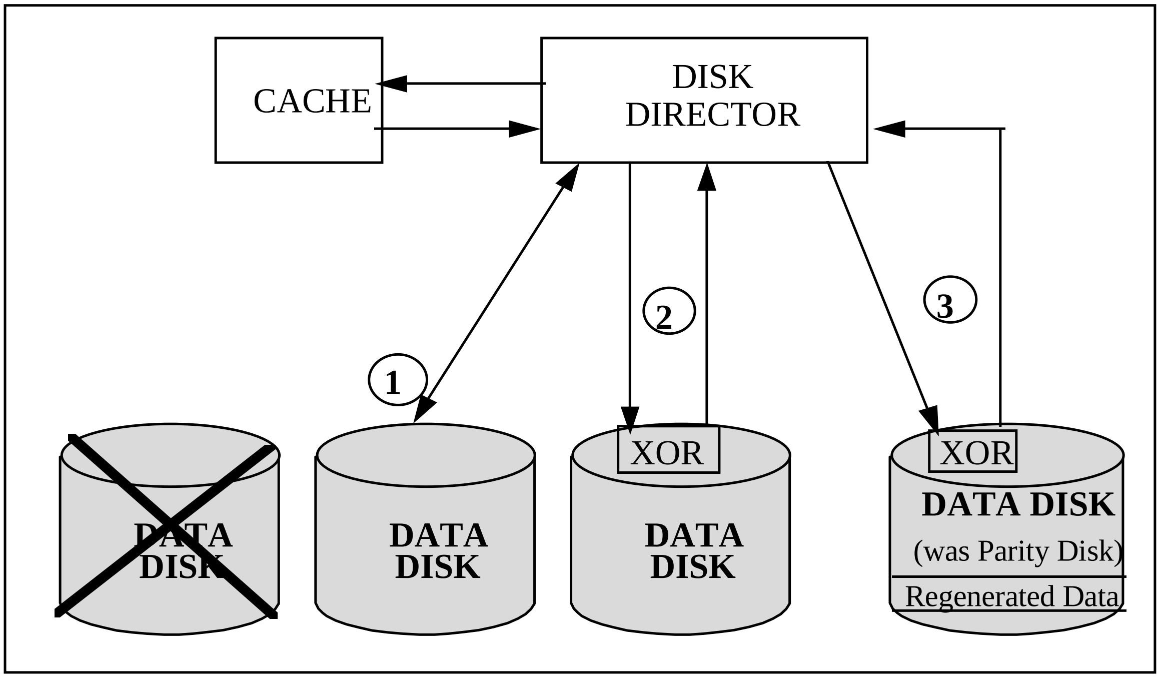

Figure 5: Regeneration Function

1	Regeneration begins with the DD issuing a standard SCSI READ command to the first surviving member in the rank and receiving the data back from the drive.
2	The data is sent to the second member using an XOR READ command with a bit set to instruct the drive to **not** write the data to the disk, but allowing it to perform the XOR computation with the data on that disk drive. The XOR'd data is sent back to the DD.
3	The DD issues another XOR READ sending the XOR'd data to the last drive in the rank (the parity drive), again with the bit set to prevent the data that was sent from being written to the disk. The data is XOR'd with the data on the disk and the result (the regenerated data) is sent back to the DD. In addition to being sent to the DD to service the request, the regenerated data is written to the parity drive as data. This improves the performance of subsequent requests for the data. The parity volume is now considered a data volume for the affected tracks.

Write operations: De-stages to the failed member of a rank first require that the data be regenerated in preparation for the write operation. The track(s) which contains the data to be written is regenerated by Boring corresponding tracks on the surviving data volumes and the parity volume. The track is then updated with the new data and written to the parity volume as data. As with read operations, this is done to improve the performance of subsequent requests for the data.

4.2.1.1 Media errors

In the event of a media error, the affected tracks will be regenerated and placed on the parity volume as data. This condition will cause the Symmetrix to place a remote service call to the Customer Support Center. The Product Support Engineer (PSE) at the support center will determine if a disk drive has been identified for replacement and dispatch a Customer Engineer. Once on site, the CE will invoke the Symmetrix Hot Replacement procedure on the service processor. The logical volumes on the disk being replaced will be placed in a not ready state and the associated ranks will begin either reduced mode or non-RAID mode of operation (depending on if the logical volume which was made not ready contained data or parity information). Once the new drive is in place, the rebuild process (described below) begins.

4.2.1.1.1 Manual Sparing

When the Symmetrix places a remote service call to report a disk drive problem, the PSE has the ability to invoke a sparing operation to a spare disk located anywhere in the system. This sparing operation will copy the data volumes from the failing disk to the spare, regenerating data where necessary to ensure a complete copy of the data volume is placed on the spare disk. While the array is in this spared state no new parity is generated, and the array operates in non-RAID mode.

When the service action is complete the data volumes will be copied to the new disk and parity will be rebuilt where necessary to return the array to normal operation.

4.2.1.2 Dynamic Sparing

The Dynamic Sparing function exploits an architectural enhancement made to Symmetrix which allows up to four copies of data to be maintained in the subsystem. As a result of this change, the minimum number of spares required to provide dynamic sparing protection for RAID-S was reduced from one per disk director to three for the entire subsystem. These spares may also be used to protect local mirrors or the local copy of a remotely mirrored pair. The sparing process itself was also improved and now works as follows:

When the Symmetrix detects the pending failure of a disk drive it establishes a mirrored relationship between the data volumes in the RAID-S group and three spare drives (which can be located anywhere in the system). The data volumes on the unaffected disks, along with readable data volumes from the failing disk, are copied to the spare disks. Data from unreadable volumes is regenerated and placed on the parity volumes of the unaffected disks as well as on the spare disks (see figure 6 below). No parity data is copied to the spare disks and no parity generation occurs since **all the data is now protected via mirroring.**

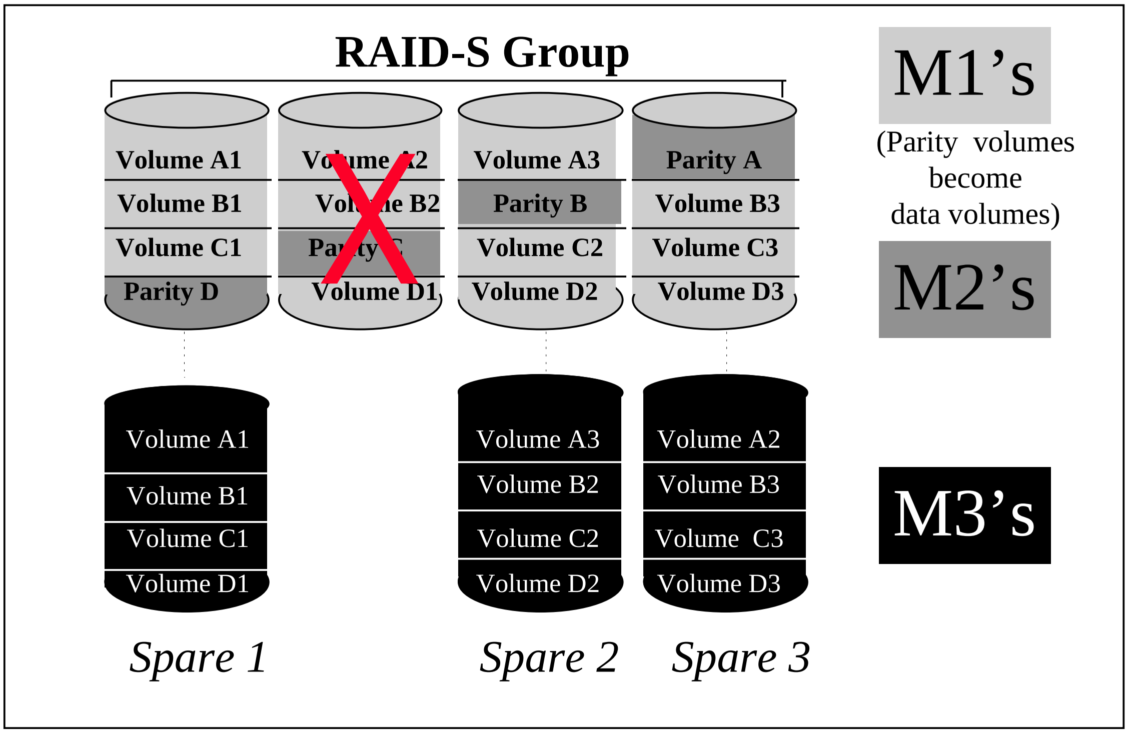

Figure 6: Dynamic Sparing

4.2.1.2.1 Partial Dynamic Sparing

In an effort to provide as much data protection as possible, the dynamic sparing process will also invoke when only one or two dynamic spares are defined or available at the time of the failure. In this case the Symmetrix will regenerate the data volumes from the failed disk and write the data to the first spare disk. Data from one unaffected disk will be copied to the second spare and mirrored relationships will be established. In this way some of the data continues to be protected from a second disk failure via mirroring.

The array will then operate as if in reduced mode. *Any* I/O to the failed volumes, or any *write* I/O to the surviving volumes, will cause a regeneration of data to the parity volume. Note that in this case data is not regenerated to the parity volumes automatically, but rather on demand. Since a full complement of spares was not available (only one or two were) full mirroring protection for the group could not be achieved, and the system will not incur the overhead of completely rebuilding the data volumes onto the parity volumes.

This will reduce the parity rebuild workload when the service action is complete and return the array to normal operation as quickly as possible

When the service action is complete and the new disk is in place, the data volumes are copied onto the replacement disk and a parity rebuild is performed for all parity volumes in the group.

This approach to full and partial dynamic sparing provides several benefits:

- Elapsed time for rebuild is lower since less reconstruction via XOR is required
- Performance during rebuild is improved through the use of mirroring.
- The amount of storage dedicated to the sparing function is less, especially in larger configurations.

4.2.2 Failed Parity Volume

The failure of a parity volume does not place a rank in reduced mode. All I/Os are serviced from the surviving data volumes as normal non-RAID requests, and parity protection is suspended for the rank. The disks then operate as normal non-RAID devices. When the parity volume is replaced, the rebuild function restores the volume as a parity volume and parity protection is resumed for the rank and RAID-S operating mode is restored.

4.2.3 Surviving Members

Read Miss Operations: Read operations to surviving members in a reduced mode rank are equivalent to non-RAID operating mode.

Write Operations: Write operations to a surviving member in a reduced mode rank triggers the regeneration of corresponding tracks for the failed member, and the writing of the regenerated data to the parity drive. Once this is complete, the write I/O is allowed to complete to the surviving member. The reason that the failed member’s data is regenerated first is because writing data directly to a surviving member would immediately invalidate the parity data. Rather than allow parity data to be invalid, the Symmetrix replaces it with valid data for the failed member, thus ensuring data integrity in the rank and improving performance both for future requests to the failed member and the resynchronization of the failed member after a service action.

Note: Gradually, the parity volume will take over for the failed data volume and service all I/O intended for the failed volume. All read and write requests to the failed volume, as well as all write requests to the surviving volumes, result in regenerated data being written to the parity volume.

5. Rebuild

When a drive in a RAID-S group is replaced, the rebuild process begins. Rebuild consists of distinct phases. The first phase is the restoration of the data volumes on the affected disk drive. This can occur in one of two ways; either regenerated data from the parity volume is copied to the data volume, or the data is regenerated from the surviving members. The second phase is the rebuilding of the parity volume, and is depicted in figure 7. During parity rebuild, only locations that stored regenerated data during the reduced mode operation are rebuilt. This helps to improve the overall rebuild time for a group.

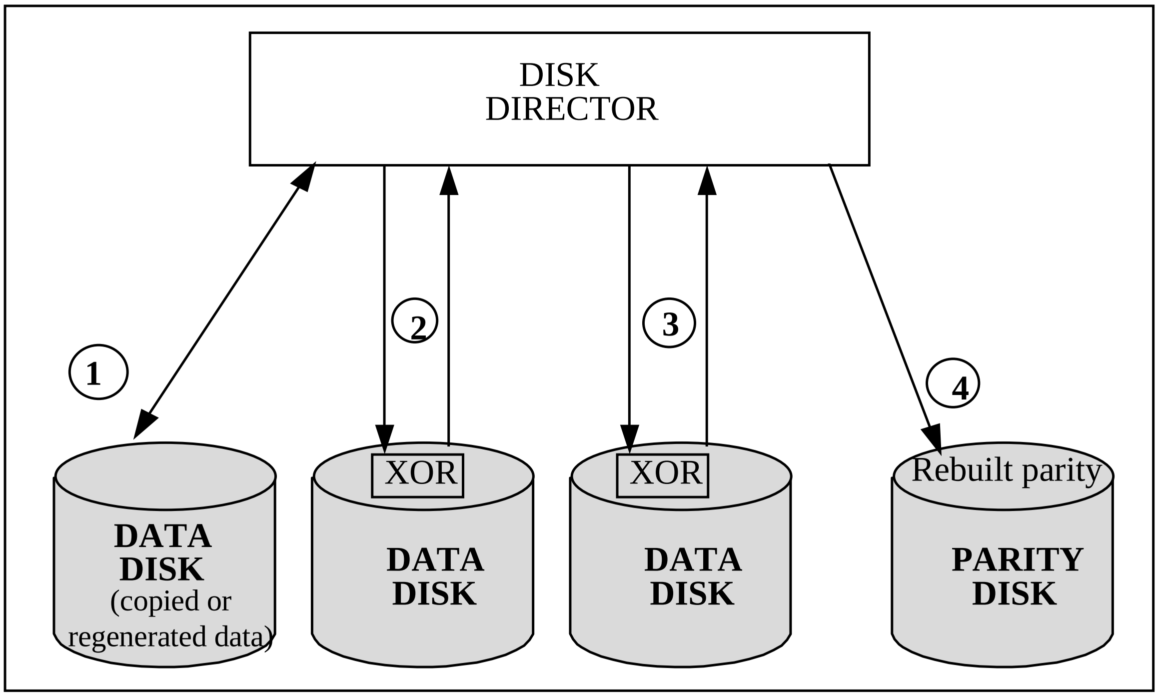

Figure 7: Rebuild Function

1	Rebuild begins with the DD issuing a standard SCSI READ command to the first data volume in the rank and receiving the data back from the drive.
2	The data is sent to the second data volume using an XOR READ command with a bit set to instruct the drive to **not** write the data to the disk, but allowing it to perform the XOR computation with the data on that disk drive. The XOR'd data is sent back to the DD.
3	The DD issues another XOR READ sending the XOR'd data to the third data volume drive in the rank again with the bit set to prevent the data that was sent from being written to the disk. The data is XOR'd with the data on the disk and the result (parity for the rank) is sent back to the DD.
4	The DD issues a standard SCSI WRITE command to write the rebuilt parity to the disk drive.

The rebuild process is invoked by a CE or PSE as part of the disk drive replacement procedure. During the rebuild process requests can continue to be serviced by the rank, however, parity protection is not restored to the rank until the rebuild operation is complete. The rebuild process is a background task that is secondary to servicing host I/O requests.

6. RAID-S Performance

6.1 RAID-S Performance Advantages

RAID-S' unique implementation and general Symmetrix architecture together offer significant performance advantages over traditional parity based RAID implementations. These advantages are summarized below:

- **Large Cache**
  Symmetrix large central cache continues to provide customers with very high read hit rates regardless of the RAID protection scheme implemented in the "back-end" of the subsystem. Cache resources are not used to store or compute redundancy data (i.e., parity or mirrored data).

- **100% Fast Write**
  The "cache all" philosophy of the Symmetrix ensures that all writes are fast writes thus ensuring the highest possible "front-end" performance for write requests. Since RAID-S made no changes to the front end of the system, the benefits of this architecture continue to accrue for RAID-S system.

- **Distributed XOR**
  The use of XOR capable disks in RAID-S improves the performance of the Read-Modify-Write sequence for parity generation compared to traditional RAID schemes. By reducing the workload on the controller, back end path contention is also reduced, contributing to faster performance when operating in normal mode.

- **Segregated RAID Groups**
  RAID-S groups are segregated from each other in the back end of the system. Rebuild activity on one group does not impact the performance of the remainder of the groups in the system.

- **Tuning**
  RAID-S keeps logical volumes intact by exploiting the Symmetrix Hyper-Volume Extension feature to map logical volumes to one and only one member disk. As a result, traditional performance tuning techniques that have been employed by storage administrators for decades can be used to tune RAID-S systems.

6.2 Performance Considerations

As with all Symmetrix ICDAs cache size and cache friendliness of the workload have a major impact on the performance delivered by the Symmetrix. Standard cache sizes for RAID-S Symmetrix systems have been adjusted upward to ensure a consistent level of performance when workloads have a heavier write orientation. As is true for all parity based RAID implementations, RAID-S is best suited to workloads whose write content is less than 25%.

The configuration flexibility of Symmetrix is an important feature in ensuring good performance for all workloads. The ability to configure a pool of RAID-1 protected volumes in a Symmetrix that is mostly protected via RAID-S is called “scalable availability”, and should be used to support applications with very high write content, such as disk to disk copies and large sequential file loading operations.

Having said that, however, it is important to keep in mind that a cache hit is a cache hit, and in this respect RAID-S Symmetrix performs in the same manner as non-RAID and mirrored Symmetrix. Understanding workload characteristics, and exploiting scalable availability where appropriate, will help ensure successful RAID-S implementations

6.2.1 Normal Mode

For read miss I/Os, RAID-S performance in normal mode is equivalent to non-RAID performance.

During periods of high utilization, I/Os may be impacted and experience a modest increase in response time. It is impossible to specifically quantify the effect since it is a function of read/write ratios, I/O rates, cache size, data blocksize, and duration of the high demand. Like other parity based RAID implementations, RAID-S does exhibit a write penalty, however the basic design of the Symmetrix (i.e. large cache) and the innovative approach taken with RAID-S minimizes the impact compared to traditional RAID implementations. Overall performance of RAID-S will obviously be dependent on the I/O rate and write content of the workload. Higher I/O rates and write content (>25%) may result in longer elapsed time due to the write penalty under these conditions.

6.2.2 Reduced Mode

Reduced mode performance of a RAID-S group is dependent upon several factors including: which volumes in the rank are being accessed, the number of volumes affected by the failure, the layout of the ranks within the Symmetrix, the I/O rate, and the read/write ratio of the workload.

Reads misses and de-stages to the failed members invoke the regeneration process for the first access to each track. De-stages to surviving members also invoke the regeneration process for the failed member (if the corresponding tracks on the failed member have not already been regenerated).

Reads misses to surviving members are treated as normal non-RAID I/Os.

These variables, the types of I/O and which volumes are being accessed, combine to make it difficult to predict the exact performance of a reduced mode RAID-S group.

6.2.3 Rebuild Mode

The performance metrics of interest in rebuild mode are *response time* for host I/Os and the *elapsed time* of the rebuild (i.e., the wall clock time spent returning the array to normal mode). These metrics are affected by the amount of host I/O, the distribution of that I/O between the replacement disk and the other disks in the array, and the read/write ratio of the workload.

Elapsed time is also directly impacted by the size of the disks in the RAID-S group. While this may seem obvious, it is often overlooked, especially when comparing different vendors RAID-5 implementations. RAID-S uses either 4GB or 9GB disks. A rebuild can clearly execute faster on a 4-GB disk since less than half the amount of data is being rebuilt. It is inaccurate to compare the rebuild times of a four disk array utilizing 4-GB drives with a four disk array which uses 9-GB drives.

It is also important to remember that the performance impact of a RAID-S rebuild is isolated to the group undergoing the rebuild. The remainder of the system is essentially unaffected.

6.2.3.1 Symmetrix RAID-S Rebuild performance

"Rebuild" is one of the three modes of operation in a RAID-S protected Symmetrix. (The other two modes are "normal" and "reduced"). A RAID-S group enters rebuild mode when a failed member of the group is replaced with a new disk and that new disk must be populated (rebuilt) with user data and parity re-calculated for the group.

6.2.3.1.1 RAID-S Rebuild Differentiators

RAID-S is implemented on a disk director level and utilizes all four SCSI buses on a disk director (i.e., a RAID-S group cannot span disk directors). **Consequently**, **the impact of rebuilding a RAID-S group is isolated to the group undergoing the rebuild and does not affect the rest of the subsystem.**

Also, RAID-S is the only parity based RAID system on the market that does not maintain parity when running in reduced mode. Rather than maintain parity when operating with a failed member, RAID-S places regenerated data on the parity volume so that subsequent requests for the same data do not incur the overhead of regeneration. This feature is exploited during rebuild mode since a new disk can be populated by copying previously regenerated data, rather than by incurring the overhead of a rebuild. This helps reduce response times for host I/O requests during the rebuild operation.

Further, RAID-S rebuild runs as a lower priority task on the disk director, so host I/O is serviced before rebuild I/O, resulting in lower response times for host I/Os.

---

[i] For a detailed description of XOR see one of the following:
"A Comparison of RAID-1 and RAID-5" *ESG Marketing Corporate SE Services* [February 13, 1995]
"What is Exclusive OR?" *SalesAdvantage* [February 27, 1995]
"The RAIDBook," *The RAID Advisory Board* [September 1, 1994]
[ii] M2 is the term used to describe the second volume in a mirrored pair.